TITLE:

格子タンパク質模型に対する新しい拡張アンサンブル法 : Multi-Self-Overlap Ensemble法(拡張アンサンブル,1998年度後期基礎物理学研究所研究会「モンテカルロ法の新展開」,研究会報告)

AUTHOR(S):

千見寺, 浄慈

CITATION:

千見寺, 浄慈. 格子タンパク質模型に対する新しい拡張アンサンブル法 : Multi-Self-Overlap Ensemble法(拡張アンサンブル,1998年度後期基礎物理学研究所研究会「モンテカルロ法の新展開」,研究会報告). 物性研究 2000, 74(2): 172-180

ISSUE DATE:

2000-05-20

URL:

http://hdl.handle.net/2433/96813

RIGHT:

京都大学学術情報リポジトリ
KURENAI 紅
Kyoto University Research Information Repository

格子タンパク質模型に対する新しい拡張アンサンブル法 — Multi-Self-Overlap Ensemble 法 —

大阪大学 理学部 物理学科 千見寺浄慈*

1 はじめに

この講演では拡張アンサンブル法の「考え方」を用いてシミュレーションが困難な系、特に格子タンパク質模型を対象にしていかに困難を解消するかを議論した。ここでの目標は平衡量や、エネルギーランドスケープがどうなっているか？ということを調べることであるので、ダイナミクスには触れない。

2 タンパク質の折り畳み問題

タンパク質は20種類のアミノ酸が1次元的につながった高分子である（つまり1次元的な情報をもっている)。タンパク質を生理条件下におくと自発的にある特定の構造（ネイティブ構造）に折り畳まれる（3次元情報)。試験管の中でも溶媒のpHを変えたりして環境をコントロールすることによってタンパク質がネイティブ構造をとったり、ランダムな構造をとることを確認することができる[1]。ここで重要なことはこの折り畳み・変性の過程が可逆的であるという事、すなわち、ネイティブ構造とは自由エネルギー最小の構造であるということである。このことから、1次元的なアミノ酸配列の情報さえあれば、原理的には3次元的なネイティブ構造が決まることがわかる。これはしばしば「第二の遺伝暗号」と呼ばれており[1]、これを解読することは現代の科学、工学の中心的課題の一つである。

さらにタンパク質の折り畳みの特徴として重要なことは通常タンパク質は10^{-3}～1秒のオーダーで折り畳むということである。一方、タンパク質のとりうる構造の数は大まかに見積もってα^Nである。ここでαは1より大きい数で、Nは鎖の長さである。この様にとりうる構造の数は天文学的な数であるにも関わらず、なぜ10^{-3}～1秒という大変短い時間でネイティブ構造を探せるか？ということをLevinthalは指摘した[2]。これを説明するために、どうやってタンパク質は全構造を探索することなく、いかに素早くネイティブ構造を捜し出すしているのか？折り畳みの物理的機構はどうなっているか？という問いに答えなければならない。これが「折り畳み問題」である。

3 格子タンパク質模型

折り畳み問題に挑戦するアプローチの仕方は様々であるが、ここではタンパク質をできるだけ簡単化して現象の本質を探るアプローチに関してのべる。

[1]第一の遺伝暗号とはDNAの塩基配列からアミノ酸配列への対応関係で、この関係については現在完全にわかっている

ここで用いる模型は「格子タンパク質模型」と呼ばれる模型で、その中でも代表的な HP モデル [3] と呼ばれる物に関して述べる。

まず、高分子鎖の構造は 2 次元正方格子上、もしくは 3 次元立法格子上の Self-Avoiding Walk で表される。アミノ酸の種類は「タンパク質の折り畳みの駆動力は疎水性相互作用が本質である」という立場に立つならばアミノ酸の種類は疎水性残基 (H: Hydrophobics) と親水性残基 (P: Polar) の 2 種類だけ考えれば良い。従ってこのモデルのアミノ酸配列は例えば $HHPPPHHPHPHP$の様に表現される。鎖の構造を$\boldsymbol{r} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_N\}$ と表したときの系のエネルギーは以下の式で定義される。

$$E(\boldsymbol{r}) = \sum_{i<j+1} u(S_i, S_j)\Delta(\boldsymbol{r}_i - \boldsymbol{r}_j) \tag{1}$$

ここで、S_i は i 番目のモノマーの種類を表す変数で $S_i \in \{H, P\}$ 、$\Delta(r_i - r_j)$ は 共有結合していない i 番目のモノマーと j番目のモノマーが最近接格子にある場合 1 をとり、それ以外は 0 をとる値で、記号 $u(S_i, S_j)$ は $u(H,H) = -1$、$u(H,P) = u(P,H) = u(P,P) = 0$ で定義される。又、この模型でのネイティブ構造は、基底状態で定義される。実際のタンパク質は（ほぼ）ユニークな構造へ折り畳まれることを考慮すると、基底状態の縮退度が少ないものがより「タンパク質らしい」とみなすことができる。特に基底状態が縮退しないものが理想的である。

HP モデルを用いることの理点は鎖の長さ Nが約 80 までなら全構造を探索することなく、しかも厳密に基底状態のエネルギー、構造、縮退度が計算できるアルゴリズム “CHCC 法 [4]” が存在するということである。つまり、HP モデルにおいては第二の遺伝暗号解読問題は解けているといえる。したがって、あとは「折り畳み機構」を理解することが重要である。

折り畳み機構を理解する為に現在必要であろうと考えられていることは「エネルギーランドスケープ」の解析である。エネルギーランドスケープを解析するためには現在のところ 2 つの困難が存在する。一つ目は、本来エネルギーランドスケープは超多次元空間であるために人間が理解しやすいように（例えば 3 次元空間で）表現する手法が未だに決定的なものがないということ[2]。二つ目は、構造探索の方法自体が難しいということである。以下ではこの後者に焦点を当てて議論する。

4 格子タンパク質模型に対するモンテカルロは難しい

現在までに良く行なわれてきた格子タンパク質模型の構造探索法は短い鎖 ($N \sim 18$) に関しては (1) 全構造厳密数え挙げ、それ以上の長さになると専ら (2) モンテカルロサンプリング法が用いられている。

鎖の長さを Nとすると全構造厳密数え挙げに関しては格子上の Self-Avoiding Walk の総数$\Omega(N)$ が$\Omega(N) \sim (N-1)^{\gamma-1}\mu^{N-1}$の様な漸近的振舞いをする[3]ので N が大きくなると$\Omega(N)$

[2]これは本研究会の「高次元位相空間の分布とダイナミクス」に関連する話題である。本稿に関連した講演として笹井氏の講演を参照。

[3]例えば 3 次元立法格子では$\mu = 4.68$、$\gamma = 7/6$ である

は天文学的な数になってしまう。従って現在の計算機の性能でも大変短い鎖しか扱うことが出来ない。著者の知る限り最も長い Self-Avoiding Walk の数え挙げは 2 次元正方格子で最大 $N = 51$[5]、3 次元立法格子で最大 $N = 18$[6] である。しかしこのような場合、メモリーを大量に使う為にあまり実用的とはいえない。現在のところ、実用的な長さは 2 次元正方格子で $N = 18$、3 次元立法格子 $N = 14$ であろう。したがってこのレベルの鎖の長さではとりうる構造の数は天文学的な数字ではなく、Levinthal が指摘したような状況にはなっていない。Levinthal が考察したような問題を議論するためには（厳密数え挙げが不可能な）もっと長い鎖を用いなければならない[4]。そのような系に対してはモンテカルロ法が有効であると期待される。しかし伝統的なメトロポリス法を HP モデルに適用すると（特に低温でシミュレートすると）望みの構造（例えばネイティブ構造や、その他の低エネルギー構造）を有限の時間でサンプルすることは事実上不可能であることがわかる [7]。いったいなぜ格子タンパク質模型に対して従来のモンテカルロ法ではうまくいかないのだろうか？

5 従来の拡張アンサンブル法

一つの原因としては、構造空間がいわゆる “Rugged energy landscape” になっているからだと考えられる。すなわち、低温で系をシミュレートすると熱的な揺らぎが小さい為にエネルギーのローカルミニマムにつかまってしまい、そこから抜け出せないという状況におちいる。このような状況を改善するための戦略として、一連の「拡張アンサンブル法」がある [8, 9, 10, 11, 12]。実際、「ローカルミニマムにつかまらない」という理点をいかしてリアリスティックなタンパク質 [13]、格子ヘテロポリマー [14]、スピングラス [11, 15] などに適用され有効性が検証されている。

6 Multi-Self-Overlap Ensemble

前章で触れたマルチカノニカル法や交換法は確かに “Rugged energy landscape” をもつシステムに対してエネルギー極小飛躍的に留まらないので、従来のメトロポリス法などと比べると、飛躍的に効率が向上した。しかし、はたしてそのまま拡張アンサンブル法を格子タンパク質模型に適用してうまくいくだろうか？実際に適用してみるとわかるが、例えばマルチカノニカル法では鎖が比較的長いとほとんどの場合、基底状態を実用的な時間でサンプルすることは不可能である。

いったいなぜうまくいかないのだろうか？問題を困難にしている原因をさらに考えると、排除体積条件がシミュレーションを困難にしているだろうと予想がつく。エネルギーが全く入っていない、高分子の模型としては最も単純な “Self-Avoiding Walk” のシミュレーションを考えてみよう。このシステムはエネルギーが入っていないので “Rugged energy landscape” にはなっていない。しかし、排除体積条件という制約がついているがために、どうしても緩和が遅くなってしまうのである [16]。つまり、いくらエネルギーのローカルミニマムに引っかからないアルゴリズムを用いてシミュレートしても排除体積条件はこれと

[4] もちろん短い鎖で議論できる側面も沢山ある。

は全く関係ないので、緩和の遅さの解決にはなっていないのである。そこで排除体積条件が問題を難しくしている原因ならばそれを緩めてやろうというアイデアが浮かぶ。ただし緩めっぱなしでは、明らかに鎖の多重占有がある構造の方がエントロピーが大きく、望みの「物理的状態」をサンプルすることができないので、「排除体積条件をゆるめはするが、積極的に緩め方を制御する」ということが必要である。

これを実現するために鎖の重なり具合を表す変数 "罰金 (V)" を導入する。罰金は以下の式で定義する。

$$V = \sum_{i \in G'} (n_i - 1)^2 \tag{2}$$

ここで G'はモノマーが1つ以上存在する格子点の集合を表し、n_iは格子点 i 上に乗っているモノマーの数を表す。結局この "罰金 V" という変数は $V = 0$ のとき、鎖は Self-Avoiding Walk が実現しており、Vの値が大きくなると多重占有が大きい "非物理的" な鎖の構造になっている、ということを表す変数である。

上で導入した "罰金" 用いて「エネルギーと罰金の2変数空間でヒストグラムが平らになる様なアンサンブル」を構成することを考える。これが実現できれば、系は局所的なエネルギー障壁に引っかからないばかりか排除体積条件に伴うトポロジカルな障壁も乗り越えられ、興味ある情報をもつ構造を効率良くサンプルできるであろう。このようなアンサンブルを "**Multi-Self-Overlap Ensemble**" と呼ぶことにする。

"Multi-Self-Overlap Ensemble" を実現する方法 [17] は容易である。まず、エネルギーと罰金の2変数空間でヒストグラムが平らになる様な重み関数 $e^{-f(E,V)}$を求める。この重み関数を求める方法は Multicanonical 法 や Entropic sampling 法と全く同じ方法で求めることができる。異なる点はヒストグラムと重み関数が2変数であるという点だけである。適切な重み $e^{-f(E,V)}$が一旦求まったらそれを用いて一回長いラン (mesurement run) を走らせる。mesurement run が吐き出したデータを用いて histogram reweighting すれば任意の温度の熱力学量が計算できる。ただし、histogram reweighting に用いるデータは Self-Avoiding Walk が実現している構造のデータだけを用いる[5]。

7 緩和の速さ – Multicanonical vs Multi-Self-Overlap –

排除体積条件を緩めたことによってどれだけ効率が良くなっただろうか？まず、"緩和の速さ" を見るために簡単な模型を用いて Multicanonical との比較を行なった [17]。ムーブは両者ともほぼ同様のものを用いた用いたモデルは2次元正方格子上、長さ $N = 54$ の HP モデルで配列は以下のとうりである。

$$P^3HP^6HP^6HP^6HP^6HP^6HP^6HP^6HP^3 \tag{3}$$

[5] したがって排除体積条件を緩めない通常のモンテカルロ等より、熱力学量の計算に使うことができるデータが減ってしまう。このことを気にする方もいらっしゃるかと思われるが、以下で見るようにその無駄を補ってあまりある以上に効率が良くなるのである。

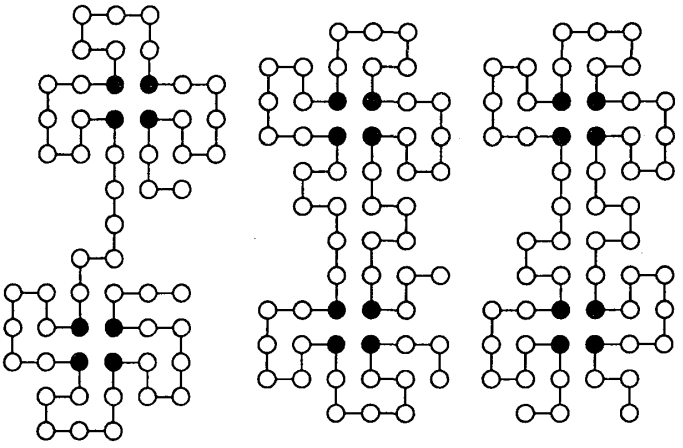


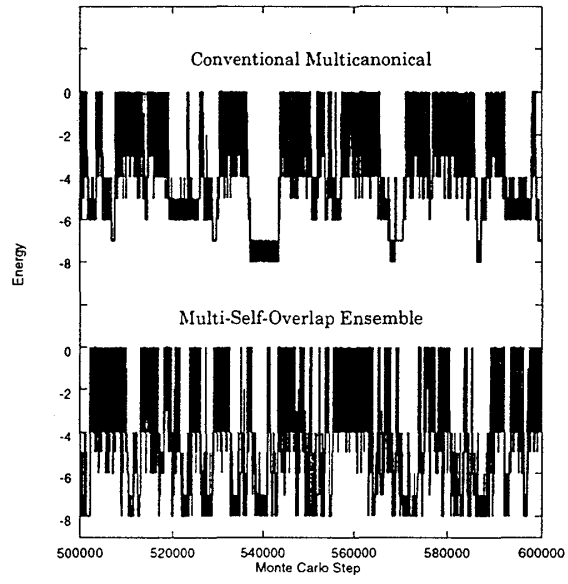
図 1: このモデルの 3 つの基底状態のタイプ：H モノマーを黒丸で、P モノマーを白丸で表している。基底状態のグループとして、左から順にグループ 1、グループ 2、グループ 3 と分類した。

この配列の基底状態のエネルギーは−8 で、H モノマーのくっつき方により図 1に示したように大きく分けて 3 種類に分類できる。これらのグループ間は通常のムーブでは[6]一旦鎖が伸び切らないと移り代わることができない。Multicanonical と Multi-Self-Overlap の両者とも十分にヒストグラムが平らになる重みを用いて mesurement run を行なって緩和を比較した。図 2(a) に両者の mesurement run におけるエネルギーの時系列を示した。Multi-Self-Overlap の方は罰金が 0 の場合のみプロットした。Multi-Self-Overlap の方が遥かに頻繁にエネルギーが上下していることから緩和が速いことを “示唆” している。さらに詳細に緩和を議論するには “すぐに構造を忘れられるか?” を見るのが必要である。そのために「基底状態をサンプルしたときに、その基底状態はどのグループか」というプロットを図 2(b) に示した。再びここでも Multi-Self-Overlap の方は罰金が 0 のときのみをプロットしている。この図から圧倒的に Multi-Self-Overlap の方が頻繁に基底状態のグループ間を移り代わっていることが分かる。つまり、本当の意味での緩和が圧倒的に速いのである。

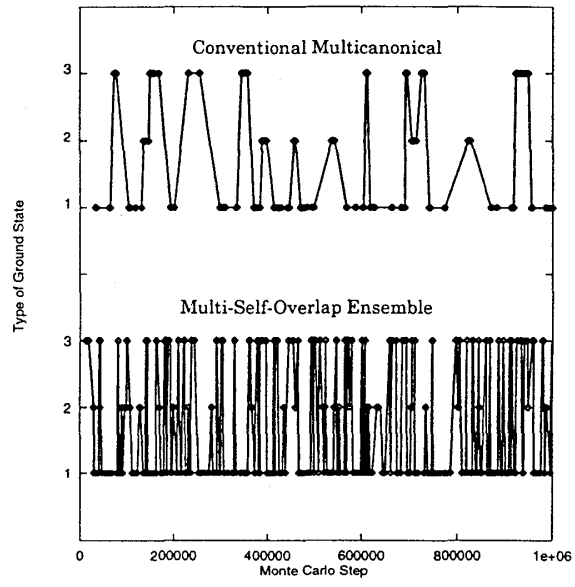
8 基底状態探索 − Multi-Self-Overlap vs その他のアルゴリズム

格子タンパク質模型の熱力学量を任意の温度で精度良く求める為には、基底状態もサンプルできることが必要条件である[7]。前節では Multi-Self-Overlap は従来の方法と比べて十分緩和が速いという事を示したが、基底状態探索としての性能はどうだろうか? もし、基底状態探索の方法としても優れているならば、格子タンパク質模型の構造探索法として非常に有

[6]ここで用いたムーブでもそうだが
[7]無論、十分条件ではない



(a) エネルギーの時系列



(b) 基底状態の移り変わり

図 2: (a) エネルギーの時系列：上に Multicanonical のものを、下に Multi-Self-Overlap での結果を示してある。(b) 基底状態の移り変わりの様子

効であるといえる。ここでは、幾つかの比較的長い HP モデルに対して Multi-Self-Overlap を用いて基底状態探索を行ない、"最適化" としての有効性を検証した [20]。さらに、その結果を現在までに開発されてきた他の基底状態探索の方法やモンテカルロ法と比較しどのアルゴリズムが最も優れているかの検討も行なわれた。いくつかの配列に対して基底状態探索を行ない、他のアルゴリズムと比較したが、ここではそのうちの一つの例を示すにとどめる。ここで示す例は 2 次元 HP モデル、鎖の長さ Nは $N = 100$ であり、配列は以下のとうりである。

$$P_3H_2P_2H_4P_2H_3(PH_2)_3H_2P_8H_6P_2H_6P_9HPH_2PH_{11}P_2H_3PH_2PHP_2HPH_3P_6H_3 \qquad (4)$$

Multi-Self-Overlap を用いて基底状態を行なった結果、約 50 時間でエネルギー $E = -50$ の構造を発見した。これを図 3に示す。この配列はすでにいくつかのグループによってそれぞれのアルゴリズムで基底状態探索が行なわれている。それらの結果と我々の結果を表 1 に記した。この配列のエネルギー $E = -50$ をもつ構造は Multi-Self-Overlap を用いて初めて発見されたものであるということを強調したい。他にもいくつかの配列で比較を行なっ

表 1: 基底状態探索の結果

Author	*Method*	*Thermal*	*E*	*Ref.*
Ramakrishnan *et al*	breaking and patching move	×	-47	[18]
Bastolla *et al*	Pruned-Enriched Rosenbluth Method	○	-49	[19]
Chikenji & Kikuchi & Iba	Multi-Self-Overlap Ensemble	○	-50	[20]

たが Multi-Self-Overlap Ensemble 法は全ての場合において最も低いエネルギー構造の発見に成功している。この結果から、現在のところ Multi-Self-Overlap が基底状態探索としても最も優れているといえる。さらに強調したいことは Multi-Self-Overlap は基底状態探索専門のアルゴリズム（単なる最適化法）ではなく熱力学量もきちんと計算できるアルゴリズムであるということである。

9 なぜうまくいったか？

ここでどうしてこんなにうまくいったか[8]を少し考えたい。まずその手始めとして、なぜ Multicanonical が ポッツモデル等で大変うまくいき、格子タンパク質では機能しなかったか?というところから始めることにする。もともと Berg らは Multicanonical を 1 次転移を起こす多状態ポッツモデルに適用した。この系は自由エネルギー障壁に起因するヒステリシスを持つために "エネルギーヒストグラムを平らにする"Multicanonical はダイレクトにこの問題を解決した。一方、スピングラスや、格子タンパク質模型に対して Multicanonical

[8] じつは著者自身もあまりにうまくいったので驚いている

を適用する際の理点は“エネルギーのローカルミニマムにとどまらない”という点であろう。しかしこれだけでは（ポッツモデルの場合と異なり）ダイレクトに困難を除去しているとはいえないのではなかろうか？[9]このことから、難しい系をシミュレートする際の“コツ”は「ある情報が欲しかったら、それをサンプルできるような“パス”を積極的に作ること」であろうと思われる。

実際 Multi-Self-Overlap では排除体積条件を緩めることによって基底状態などの重要な情報をもった構造へのパスを作っている。これを図4に示す。これは基底状態のエネルギー $E=-42$ を持つ配列に関して実際にシミュレーション中で実現された (E,V) 平面上での遷移を線で結んだものである。この図から、基底状態などの興味ある状態、すなわち Self-Avoiding Walk が実現し、かつ低いエネルギー状態というのは、Self-Overlap がある構造とつながっていることがわかる。例えば $(-42,0)$ と $(-41,0)$ の間の遷移が見られない様に、Self-Avoiding Walk 間の遷移は見られないことに注目されたい[10]。

一見すると Multi-Self-Overlap Ensemble は鎖の多重占有があるという非物理的状態が多数出現するために不自然なことしているかのように見えるが、実は「排除体積条件を緩めることによって低エネルギー状態へのパスを作った」という意味で格子タンパク質模型に対するアプローチとしては従来の拡張アンサンブル法よりもはるかに自然な発想なのである。

以上の結果を考慮すると、これからの拡張アンサンブル法は困難の原因となっているものを除去する方向[11]へ拡張するのがより発展性があるのではないかと思われる。もちろん、どんな場合でも拡張する方向が容易にわかるわけではなく、それを選ぶところで物理的センスが問われるかもしれない。

10 おわりに

以上、高分子鎖の排除体積条件を緩める方向に拡張したモンテカルロ法– Multi-Self-Overlap Ensemble –の格子タンパク質模型への適用例を紹介してきた。くりかえしになるが、これからの拡張アンサンブル法は困難の原因を見極めてそれを除去する方向に拡張することが重要であろうということを強調したい。最後に本研究の共同研究者である菊池誠氏、伊庭幸人氏に感謝したい。

参考文献

[1] A.C. Anfinsen: Science **181**, 223 (1973)

[2] C. Levinthal: J. Chim. Phys. **65**, 44 (1968)

[9]無論、従来のメトロポリス法などよりは遥かに良いが

[10]この図では省略されているが、もちろんエネルギーの高いところでは Self-Avoiding Walk 間の遷移も見受けられた。

[11]例えその方向が非物理的な方向だとしても

[3] K.F. Lau, K.A. Dill: Macromolecules **22**, 3986 (1989)

[4] K. Yue and K.A.Dill: Phys. Rev. E. **48**, 2267 (1993)

[5] A.R. Conway and A.J. Guttmann: Phys. Rev. Lett. **77**, 5284 (1996)

[6] V.S. Pande, A.Y.Grosberg and T. Tanaka: J. Chem. Phys. **107**, 5118 (1997)

[7] K. Yue, K.M. Fiebig, P.D. Thomas, H.S. Chan, E.I. Shakhnovich and K.A. Dill: Proc. Natl. Acad. Sci. USA **92**, 325 (1995)

[8] B.A. Berg and T. Neuhaus: Phys. Lett. **B 267**, 249 (1991)

[9] B.A. Berg and T. Neuhaus: Phys. Rev. Lett. **68**, 9 (1992)

[10] J. Lee: Phys. Rev. Lett. **71**, 211 (1993)

[11] K. Hukushima and K. Nemoto: J. Phys. Soc. Jpn. **65**, 1665 (1996)

[12] E. Marinai and G. Parisi: Europhys. Lett. **19**, 451 (1992)

[13] U.H.E. Hansmann and Y. Okamoto: J. Comp. Chem. **14**, 1333 (1993)

[14] N. Urakami and M. Takasu: J. Phys. Soc. Jpn **65** 2694 (1996)

[15] B.A. Berg and T. Celik: Phys. Rev. Lett. **69**, 2292 (1992)

[16] A.D. Sokal: Monte Carlo methods for the Self-Avoiding Walk
in *Monte Carlo and Molecular Dynamics Simulation in Polymer Science*, ed. K. Binder (Oxford Unversity Press, New York, Oxford , 1995) 47.

[17] Y. Iba, G. Chikenji, and M. Kikuchi: J. Phys. Soc. Jpn. **67**, 3327 (1998)

[18] R. Ramakrishnan, B. Ramachandran, and J.F. Pekny: J. Chem. Phys. **106**, 2418 (1997)

[19] U. Bastolla, H. Frauenkron, E. Gerstner, P. Grassberger and W. Nadler: Proteins: Struct.Funct.Genet.**32**, 52 (1998)

[20] G. Chikenji, M. Kikuchi, and Y. Iba : cond-mat/9903003